Kevin Miele

3/20/24

DATA 110

Prof. Muhammad

Project 1

Pokémon and The Pokémon Go Dataset

Pokémon is a Japanese media franchise that takes place in a universe where humans co-exist with creatures called Pokémon. In this universe, humans can collect and train Pokémon to fight against each other. Each Pokémon has a unique set of powers. Many products have been created from this franchise, and the three most popular are trading card games, video games, and TV shows. Pokémon has always been a part of my childhood ever since I started elementary school. Me and my friends would trade (or steal) each other’s cards and we would always compete to see who had the best collection. The first time Pokémon was introduced to me as a kid was when I saw a movie of Pokémon: Seaside Pikachu on VHS. It opened up the world of Pokémon for me, so the series, games, and cards hold a special place in my heart. That is why I decided to choose this dataset.

This dataset (pokemon.csv) is from the popular mobile gaming app Pokémon Go. I made this conclusion after seeing that Pokémon Go was released on July 6th, 2016. The dates where these Pokémon were captured in the dataset were all within December 2016. The days the Pokémon captures were recorded were the 13th through the 18th. Pokémon Go is a location-based game where users can use their phones to capture Pokémon in real-time through their camera. Pokémon Go uses popular tourist sights like for example: The Eiffel Tower as a “Pokémon Gym” where trainers (in the game users are labeled as Pokémon trainers) can battle rival teams to gain control over that specific gym. That is why there are many different Trainer regions and subregions listed in the dataset because Pokémon Go is available worldwide. The dataset even displays the Pokémon's region showing where they are originally from. It uses the GPS that almost all up-to-date mobile phones have.

The dates and times are pretty self-explanatory, and the third row displays the unique names of the Pokémon. The fourth, fifth, and 6th rows I explained above when describing regions and sub-regions. Here is where more of the data gets interesting: For Level: Whichever number is under the Level row is the base level the Pokémon has when it is captured. For Level Met: Level met is the level that you can upgrade the Pokémon to. It’s kind of like a level goal. So, using the first column as an example, the player surpassed the Level met goal and managed to increase their Pokémon’s level (Oricorio) to 13. For the Gender row, Pokémon can either be Male, Female, Or Neutral.

For the Type 1 and Type 2 rows: Types are aspects of Pokémon and attacks. Each Pokémon can have one or two types. When it has only one type, it’s called a pure-type Pokémon. When it has two types, it’s called a dual-type Pokémon. The different types are Bug, Dark, Dragon, Electric, Fairy, Fighting, Fire, Flying, Ghost, Grass, Ground, Ice, Normal, Poison, Psychic, Rock, Steel, and Water.

For the Nature row: Natures are a Pokémon’s personality and can affect its stats and leveling up. The Poke Ball row: There are many Poke Balls in the game: Poke Ball, Great Ball, Ultra Ball, Master Ball, Premier Ball, and Beast Ball. The power levels they possess go from low to high starting with Poke Ball and ending with Master Ball. There are other Poke Balls that stand out on their own like the Premier Ball, Beast Ball, Dusk Ball, Quick Ball, Heal Ball, and more. The ones that stand out are more likely for special challenges. For this dataset, it will be easier and more organized to focus on the main 4.

The Held Item Row: This is when a player uses items like boosts and powerups to have a better chance of damaging and capturing a Pokémon. I assume that T stands for True, and F stands for False. So, T(True): A player used items to capture. And F(False): A player did not use items to capture.

The last row: Perfect IV’s: IV stands for Individual Value. A Pokémon has 3: Attack, Defense, and Stamina. Each IV is out of 15. So a Perfect IV would be 15/15.

One of the first trends that immediately stood out to me was the Pokémon types. The most common type of Pokémon that trainers caught were: Flying. Here is a more detailed analysis of the common types:

Top 5 most common types of Pokémon:

Flying 97.0

Normal 93.0

Water 88.0

Dark 48.0

Bug 46.0

To find this data, I first loaded the dataset from Google Drive to my GitHub account, and then in Google Colab, I started using the pandas library (a powerful tool for data manipulation and analysis in the Python programming language). Then, I extracted the frequency of each type from both the "Type1" and "Type2" columns, as Pokémon can have up to two types. I combined the counts from both columns to get a comprehensive view of the overall type distribution.

To visualize the results, I created a bar chart using the matplotlib library (a cross-platform data visualization and graphical plotting library for Python). The chart displayed the top N (Number of Top Types) most common types of Pokémon, where N was set to 5 in this case. The bar chart provided a clear and concise representation of the frequency of each type, making it easy to identify the most prevalent Pokémon types at a glance.

Another part of the dataset I wanted to analyze was the Trainer regions that were associated with each capture. I wanted to find out which regions were the most popular and I wanted answers to if regions influenced the amount of Pokémon that a player encounters. I again loaded the Pokémon dataset from my GitHub link using the pandas library. I counted the number of captures for each Trainer Region using the value\_counts() function, which provided a count of Pokémon captures for each region. After that, I created a pivot table from the counted data, pivoting on Trainer Region to form the rows and Captures to form the columns, organizing the data for heatmap visualization. I filled any missing values in the pivot table with zeros to ensure all regions were represented in the heatmap. For that specific process, I ran into a lot of errors, but my main goal was to display the zeros so that the other values could stand out from the rest. I thought it looked like a cool Matrix-style visualization. Finally, I plotted the heatmap using seaborn's heatmap() function (seaborn is another data visualization library based on matplotlib), specifying the colormap and formatting for clarity and visual appeal. This process allowed for a clear visualization of the popularity of Trainer Regions in the dataset based on Pokémon captures.

Pokémon Go relies heavily on its player base, and many of the game features like Poke-Stops and Gyms are based on historical landmarks, art installations, and areas of interest. This is going back to the Pokémon Gym and the Eiffel Tower I mentioned earlier. Looking at the scatterplot and even forming a general consensus without it, you can conclude that most of the player base and activity would reside in Japan and the United States considering it was the most popular in those countries. In 2016, it was the biggest mobile app game at the time. This is an important detail to consider because as the player base increases so does the more Pokémon that are likely to spawn as well as more Poke stops and Gyms. The way the game was kept alive was through popularity.

For the last visualization, I wanted to make a scatterplot on the Pokémon types. I used a bar graph first, but I wanted to dive deeper considering that some types are paired together and have more details to discover. The Pokémon types in Pokémon Go work and are displayed interestingly, just like the card game. I wanted to make it as appealing as possible because each Pokémon type has a unique color. The scatterplot shows which types were the most common and which types matched the most. I included a grid in the background so that the viewer can easily match the types together from the x and y-axis. I used seaborn's scatterplot function once more to create the scatterplot, specifying the dataset, x-axis attribute (Type1), y-axis attribute (Type2), and creating hues to differentiate colors by each type. I adjusted the size of the markers to enhance visibility and included a legend to help interpret the colors corresponding to each Pokémon type.

This dataset can help us conclude that there are many factors when it comes to creating a mobile game and even a franchise in general, especially location-based. With popularity, many companies must consider where they release their product in the world and how it will resonate with the average person. Pokémon has been around since the 90s and became popular way before Pokémon GO, so the franchise already knew their fanbase and used this to their advantage. You can especially see this in the way the data is represented in the different regions.

When you also look at the bar graph and scatterplot of the Pokémon types, you can see how well certain types interact with each other and which ones are the most popular/common. The reason why Pokémon is so popular and special to people is because Pokémon and Pokémon types can be connected to a human’s personality. If someone is interested in astrology or the supernatural, they would most likely capture more Dark, Ghost, and Psychic types of Pokémon. If someone is interested in action or action movies, they would gravitate towards Fire, Electric, and Dragon Pokémon. This then ties into the Nature of each Pokémon which could be explored more in the dataset. For example, matching a person’s personality type and the Pokémon they are the most likely to capture. That is what makes this dataset unique because it can be interpreted in many different ways, and it can be paired with many other datasets. This is a small sample of data for a game that has over 1 billion downloads, so you can imagine the possibilities when it comes to bigger and bigger datasets. The main takeaway I got from this dataset is that Pokémon and Pokémon GO as a franchise is a way for everyone around the world to connect. Growing up we all had a cartoon character or figure in our life that influenced us in some way, and I think that’s what Pokémon aspires to do when it comes to connection and being human. It also allows us to look at the data from a business perspective. When creating a franchise, what would I want most people to connect with and how? Which countries and cities would I think my franchise would thrive in? It’s these perspectives that allow us to go beyond the data and develop different ideas and creativity.